RAFDA of PAKDD

Sentiment Analysis beyond Public Opinions

Yi Chen

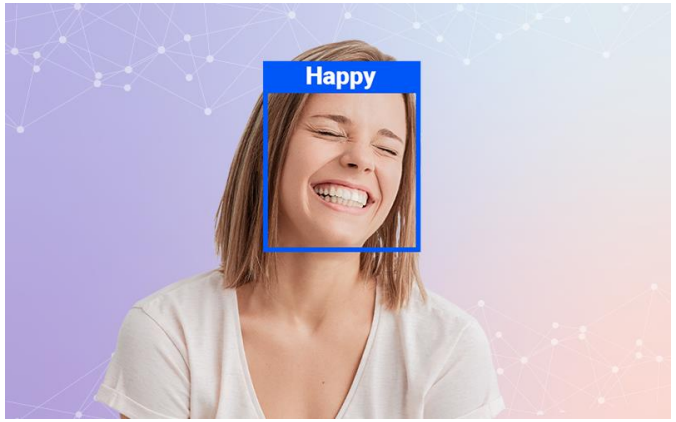
New Jersey Institute of Technology

In collaboration with Weiting Gao, Xiangyu Gao, Wenjin Chen, Dr. David J. Foran

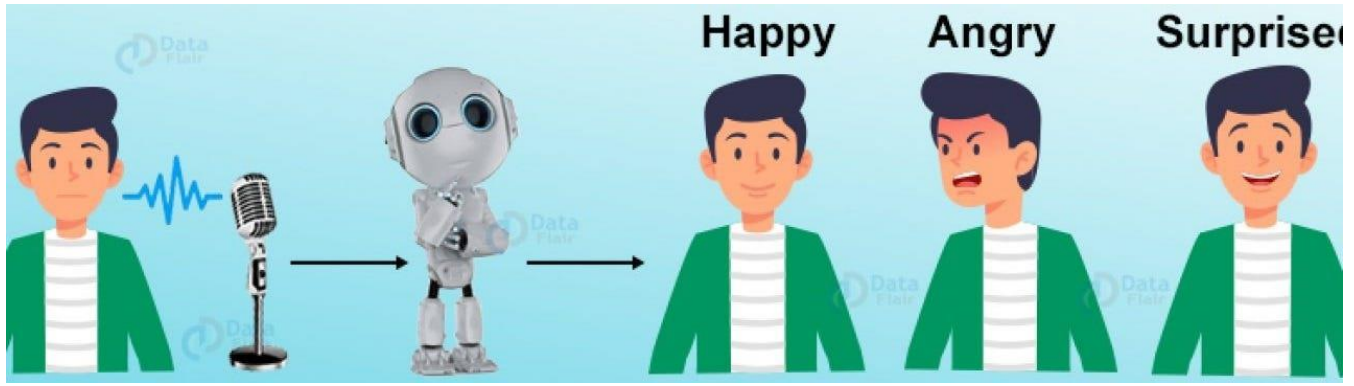
Funded by NIH, the Martin Tuchman'62 Chair Endowment, and the Leir Foundation

Affective Computing

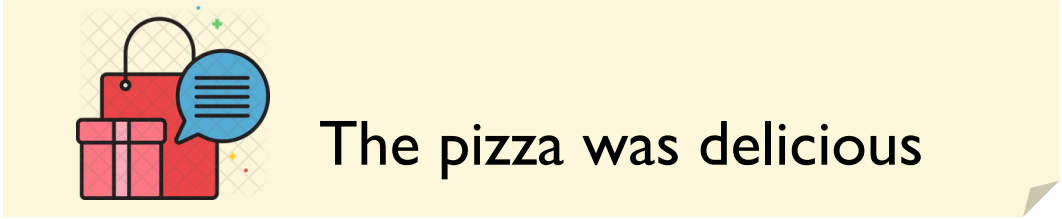
Affective Computing: the development of technologies that can recognize, interpret, process, and simulate human emotions.



Visual Emotion Recognition



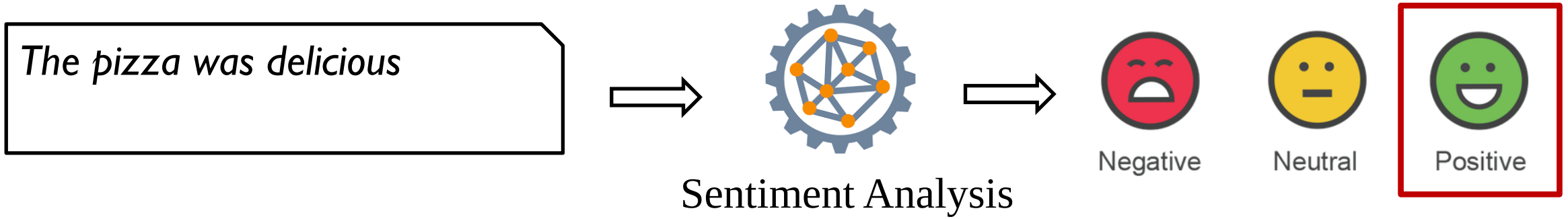
Speech Emotion Recognition



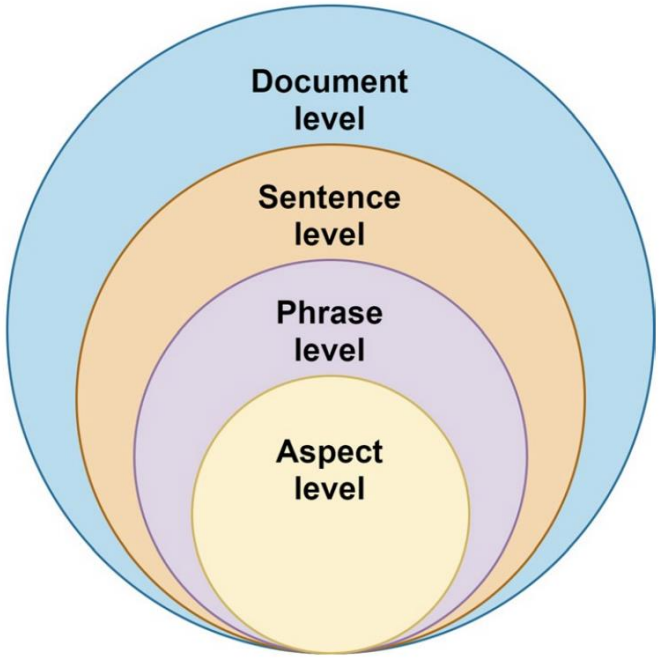
Sentiment Analysis

Sentiment Analysis

Sentiment Analysis: NLP techniques that analyze text to determine the emotion tone to be positive, negative or neutral



Aspect-Based Sentiment Analysis (ABSA)



Aspect-Based Sentiment Analysis (ABSA):
NLP techniques that identifies emotions to different aspects of a product or service.

The pizza was delicious, but the waiter was rude

ABSA Techniques

Aspect	Polarity
Pizza	positive
Waiter	negative

Applications: Mining Public Opinions

Product Reviews

Bipin R.

★★★★★ **Nice, Simple to install and use and effective!**

Reviewed in India on March 12, 2022

Color: Venetian Bronze | Configuration: Doorbell only | **Verified Purchase**

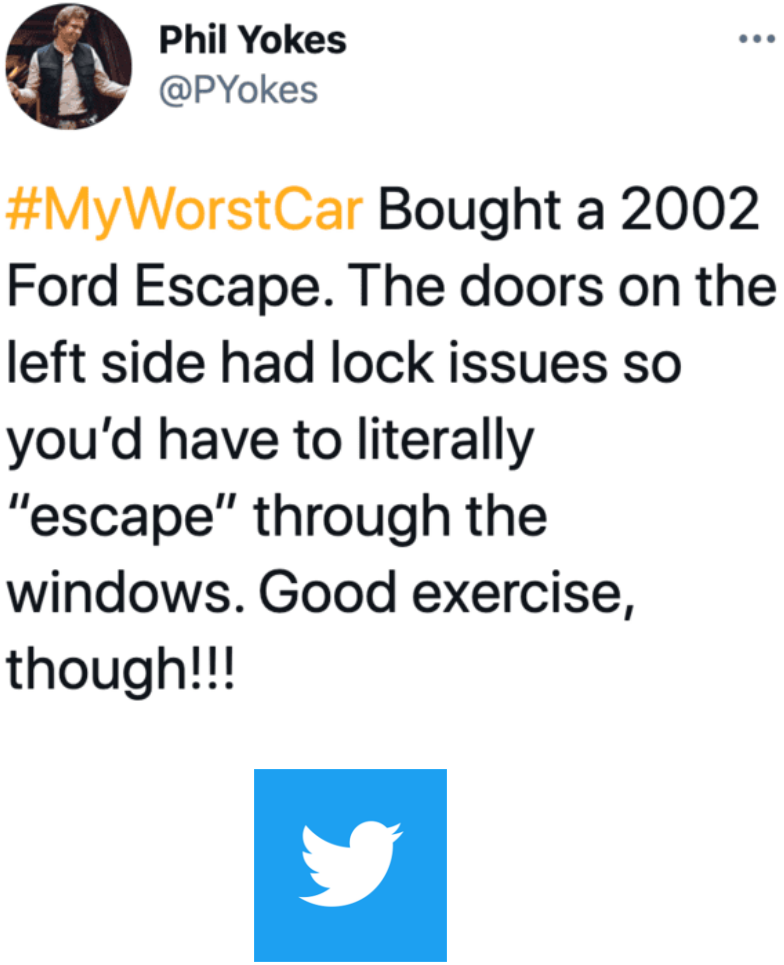
Works well. Bell rings in our phones and now all can see who's at the door from anywhere in the world and not just at home

2 people found this helpful

amazon

yelp

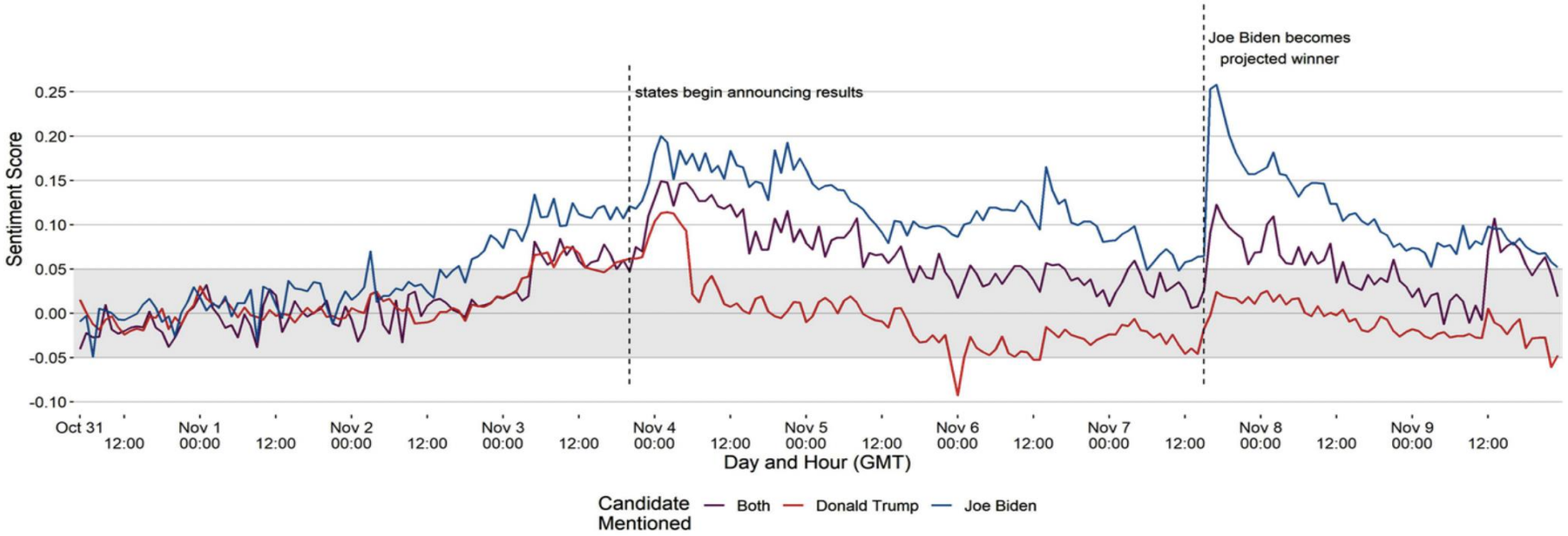
Applications: Mining Public Opinions



Product Reviews

Brand Reputation

Applications: Mining Public Opinions



Product Reviews

Brand Reputation

Election Forecast

Applications: Beyond Mining Public Opinions

Predict

Stock Price

Finance

Exchange Rate

Market Trend

Applications: Beyond Mining Public Opinions

Finance

Predict

Legal

Judge's Decision

Applications: Beyond Mining Public Opinions

Predict

Clinical Outcome,
e.g. in-hospital mortality

Finance

Legal

Medical

Outline

- Introduction of SA
- **A case study of SA in medical domain**
  - **Cancer biomarker information extraction from pathology reports [PAKDD'24]**
- Discussions and Future Directions

Background: Biomarker Information

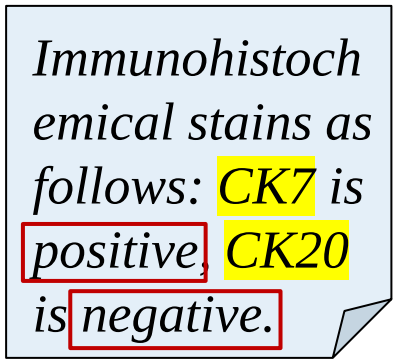
- **Cancer biomarkers**: molecules (e.g., genes, DNA, proteins) that indicate the risk of cancer.
- **Results of cancer biomarkers**: positive, negative, or unknown
- Biomarker information is used for:
  - Assess the risk of cancer
  - Monitor a disease's progression
  - Check the effectiveness of a therapy

  …

Example of pathology report:

Immunohistochemical stains as follows:
CK7 is positive, CK20 is negative.

Biomarker Information Extraction (BIE)



Automatically

Unstructured data

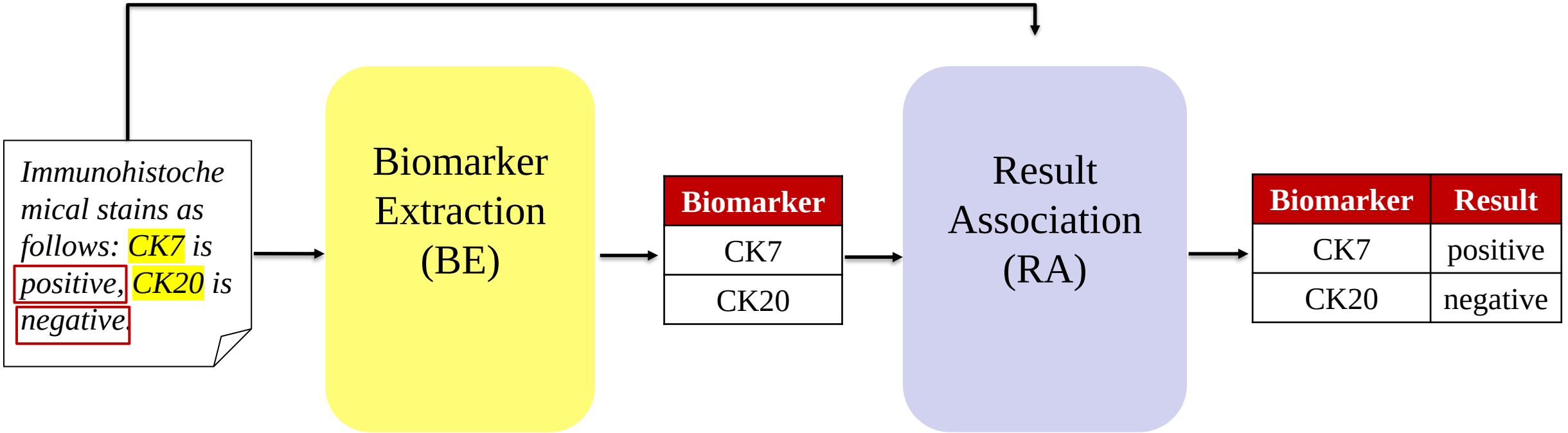
Biomarker	Result
CK7	positive
CK20	negative

Structured data

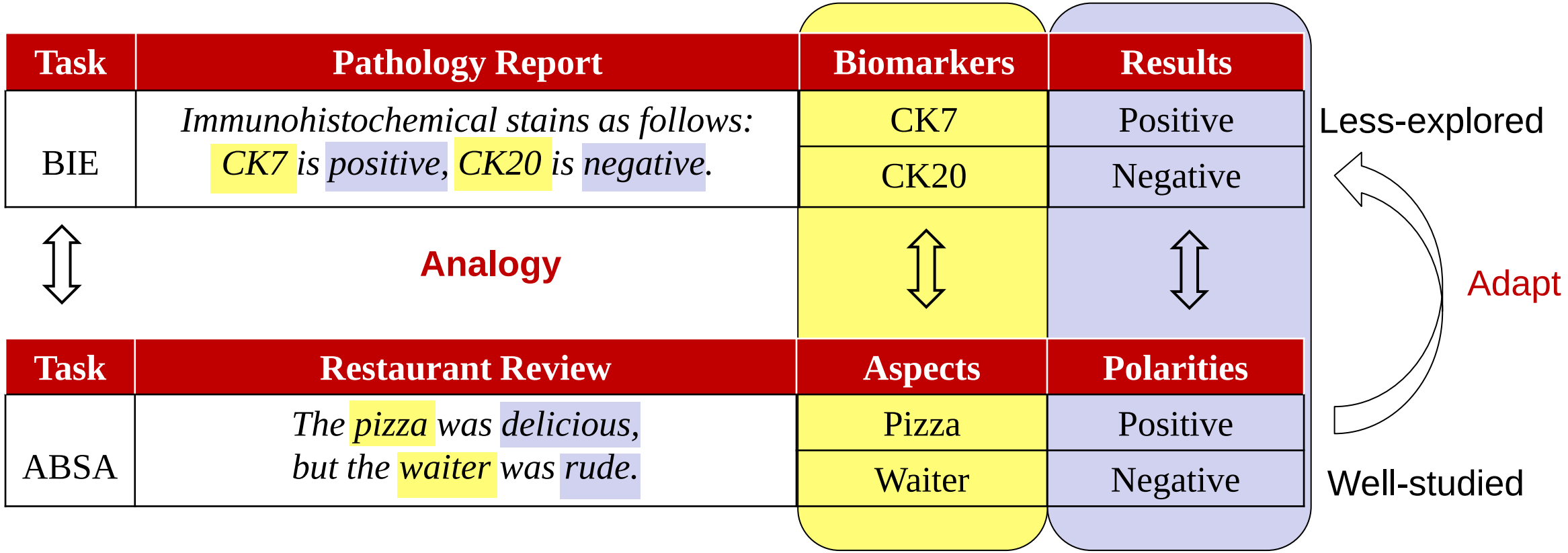
Downstream tasks, e.g.,

- correlation of biomarkers and cancers,
- cancer risk prediction

BioReX (**Bio**marker and **Re**sult E**X**traction Model)



Analogy with Aspect-based Sentiment Analysis (ABSA)



Challenges for Biomarker Information Extraction

- Mention Overlapping (BE): same mentions have different meanings

Case 1: *Part E is labeled ... right posterior X, EO left anterior X, EP EQ left posterior X, ER right anterior Z, ES right mid Z ...*
Case 2:*The tumor ... is negative for ER, PR, RCC, CA19.9, PAX8, and Vimentin.*
Case 3:*... patient who presented to the ER at [location] on [date] with abdominal pain…*

- Mix-polarity (RA): a sentence contains multiple biomarkers with different results

Case 4. *Immunohistochemical stains as follows: CK7 is positive, CK20 is negative, WT1 is positive, PAX8 is positive, S100P is negative, GATA3 is negative*

- No-result Cases (RA): some biomarkers do not have results

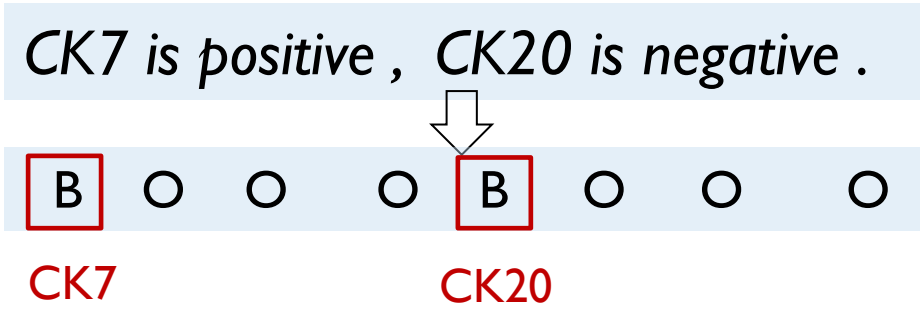
Case 5. *Note: ER = Negative, PR = Negative, HER2 by fish is pending and the result will be given in an addendum report*

Adapt ABSA Techniques for BIE

Challenges	Adapting		Solutions
	ABSA	BIE	
Mention Overlapping (BE)	limited labeled data General domain →	Medical domain	BioBERT post-training Enhance cancer knowledge
Mix-polarity (RA)	Less intensive →	More intensive	Two attentions - syntactic-based - semantic-based Adjust the contribution of context words
No-result (RA)	X →	No-result cases are common	

Biomarker Extraction (BE)

- Goal: Extract biomarkers
- A sequence labeling task: label each token as **B**egin, **I**nside, or **O**utside



- Leverage BioBERT and post-training to enhance the model on capturing the context of biomarkers

BioBERT

✔ medical domain knowledge
✔ context information
X cancer knowledge
X task knowledge

Post-Training by Medical Term Extraction

✔ external data source
✔ cancer knowledge
✔ task knowledge

BioX

Used for fine-tuning, i.e., sequence labeling for BE

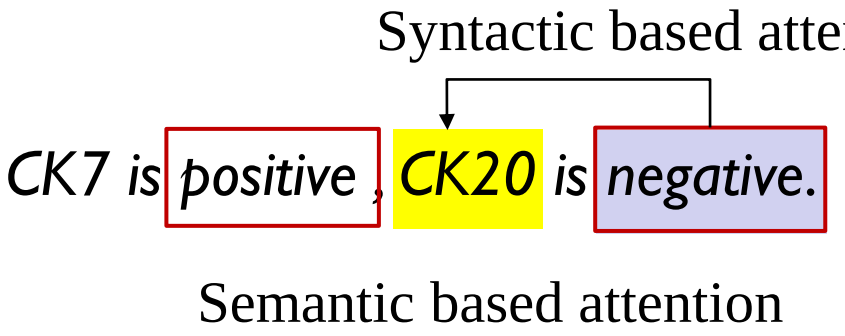
Result Association (RA)

- Goal: identify the result of the extracted biomarkers as *Positive*, *Negative or None*.

Sentence: *CK7 is positive , CK20 is negative .*
Biomarker: *CK20*

⇨ **Negative**

- The correct result is expected to have largest contribution to the target biomarker.
  - Adjust the contribution of context words to the target biomarker by two attention modules, Semantic based & Syntactic based



Experimental Settings

Dataset	Details	Used for
Pathology reports from Rutgers Cancer Institute of New Jersey	995 reports 43,423 annotated sentences	End-task
MedMentions dataset	4,000 biomedical abstracts, 47,722 sentences, 321,899 annotated entities. (20% cancer related)	Post-training

Experimental Results

Methods		Biomarker Extraction			Result Association		
		Precision	Recall	F1	Precision	Recall	F1
Baselines	cTAKES	0.622	0.144	0.224	-	-	-
	CLAMP	0.283	0.374	0.323	-	-	-
	CBEx	0.615	0.884	0.725	-	-	-
	BioBERT	0.916	0.922	0.918	0.927	0.919	0.920
Proposed Models	BioX	0.934	0.937	0.935	0.929	0.922	0.923
	BioX + SynAtt	-	-	-	0.945	0.940	0.946
	BioX + SemAtt	-	-	-	0.939	0.932	0.934
	BioReX	-	-	-	0.954	0.950	0.952

- Clinical NLP software have inferior performance than DL methods
- Post-Training is effective.
- BioX performs best.

- BioBERT and BioX have similar performance.
- Each attention module introduces performance gain.
- BioReX performs best.

Case Studies: Mention Overlapping

Ground Truth

Case 1: Part E is labeled ... right posterior X, EO left anterior X, EP EQ left posterior X, *ER* right anterior Z, ES right mid Z ... — False

Case 2:*The tumor ... is negative for ER, PR, RCC, CA19.9, PAX8, and Vimentin.* — True

Case 3:*... patient who presented to the ER at [location] on [date] with abdominal pain…* — False

Biomarker Extraction ⇨

✓ : Correct extraction
X : Incorrect extraction

	Case 1	Case 2	Case 3
cTAKES	✓	X	✓
CLAMP	✓	X	✓
CBEx	✓	✓	X
BioBERT	✓	✓	X
BioX	✓	✓	✓

Case Studies: Mix-Polarity and No-Result

Case 4. *Immunohistochemical stains as follows: CK7 is positive, CK20 is negative, WT1 is positive, PAX8 is positive, S100P is negative, GATA3 is negative*

Case 5. *Note: ER = Negative, PR = Negative, HER2 by fish is pending and the result will be given in an addendum report*

Ground Truth: biomarkers with positive, negative or no results are shown in red, blue, brown, respectively.

Result Association ⇨

✓: Correct association
X: Incorrect association

	Case 4 (mix-polarity)						Case 5 (no result)		
	CK7	CK20	WT1	PAX8	S100P	GATA3	ER	PR	HER2
BioBERT	✓	X	X	X	X	✓	✓	✓	X
BioX	✓	X	X	✓	X	✓	✓	✓	X
BioReX	✓	✓	✓	✓	✓	✓	✓	✓	✓

Summary

- We propose an analogy between BIE and ABSA, opening pathways for the less-explored BIE domain to leverage the techniques in the well-studied ABSA field.
- The proposed method, BioReX, addresses the unique challenges in BIE.
  - Mention overlapping: post-train BioBERT to enhance it with domain and task knowledge
  - Mix-polarity and no-result: syntactic-based and semantic-based attention modules
- The first study that extracts both cancer biomarkers and their results in a limited data setting.
- We are in the process of deploying BioReX in clinical data warehouses.

Outline

- Introduction of SA
- A case study of SA in medical domain
  - Cancer biomarker information extraction from pathology reports
- **Discussions and Future Directions**

Opportunities in Mining "Opinions" of Professionals

	Documents	Opinions	Sentiment
Financial News	*The acquisition will considerably increase Kemira 's sales and market position in the Russian metal industry coatings market.*	assess the impact of acquisition	
Legal Docs	*Here, the facts do not support a finding of jurisdiction for the present dispute, the in rem action against the Defendant Vessel is struck out*	judge the facts	
Clinical Notes	*He was treated with both Zosyn and linezolid in the MICU with improvement in his clinical status. Urinary tract infection has resolved.*	evaluate patients' medical condition	

What other domains SA can be applicable?

Challenges in Mining "Opinions" of Professionals

Domain-specific

Long and complex

Fine-grained analysis, ABSA, is often necessary

E.g., The acquisition will considerably increase Kemira 's sales and market position.

Aspect	Polarity
sales	positive
market position	positive

Take-home

- SA beyond mining public opinions
- Opportunities in mining professional documents
- Challenges
  - Complex documents
  - Aspect-based SA holds potential
  - Work with domain experts

Thank you!

Questions?

References

[1] https://en.wikipedia.org/wiki/Affective_computing

[2] Dhillon, A., Singh, A., Bhalla, V.K.: A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: from computational needs to machine learning and deep learning. Archives of Computational Methods in Engineering 30(2), 917–949 (2023)

[3] Liu, H., Chatterjee, I., Zhou, M., Lu, X.S., Abusorrah, A.: Aspect-based sentiment analysis: A survey of deep learning methods. IEEE Transactions on Computational Social Systems 7(6), 1358–1375 (2020)

[4] Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., Xu, H.: CLAMP– a toolkit for efficiently building customized clinical natural language processing pipelines. JAMIA 25(3), 331–336 (2018)

[5] Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. JAMIA 17(5), 507–513 (2010)

[6] Islam, M.T., Shaikh, M., Nayak, A., Ranganathan, S.: Extracting biomarker in- formation applying natural language processing and machine learning. In: ICBBE. pp. 1–4. IEEE (2010)

[7] Lee, J., Song, H.J., Yoon, E., Park, S.B., Park, S.H., Seo, J.W., Park, P., Choi, J.: Automated extraction of biomarker information from pathology reports. BMC medical informatics and decision making 18(1), 1–11 (2018)

References

[8] Gao, X., Shi, J., Chen, W., Sazo, N., Chu, H., Sadimin, E., Foran, D.J., Chen, Y.: CBEx: A hybrid approach for cancer biomarker extraction. In: BIBM. pp. 2958–2958. IEEE (2020)

[9] Zhang, X., Zhang, Y., Zhang, Q., Ren, Y., Qiu, T., Ma, J., Sun, Q.: Extracting comprehensive clinical information for breast cancer using deep learning methods. International Journal of Medical Informatics 132, 103985 (2019)

[10] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), 1234–1240 (2020)

[11] Xu, H., Liu, B., Shu, L., Philip, S.Y.: BERT post-training for review reading com- prehension and aspect-based sentiment analysis. In: ACL. pp. 2324–2335 (2019)

[12] Liu, Q., Zhang, H., Zeng, Y., Huang, Z., Wu, Z.: Content attention model for aspect based sentiment analysis. In: WWW. pp. 1023–1032 (2018)

[13] Mohan, S., Li, D.: MedMentions: A large biomedical corpus annotated with UMLS concepts. In: 1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019 (2019). https://doi.org/10.24432/C5G59C, https://doi.org/10.24432/C5G59C

[14] M.-Y. Day and C.-C. Lee, ‘‘Deep learning for financial sentiment analysis on finance news providers,’’ in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2016, pp. 1127–1134.

References

[15] W. Souma, I. Vodenska, and H. Aoyama, ‘‘Enhanced news sentiment analysis using deep learning methods,’’ J. Comput. Social Sci., vol. 2, no. 1, pp. 33–46, Jan. 2019.

[16] S. F. Crone and C. Koeppel, ‘‘Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons,’’ in Proc. IEEE Conf. Comput. Intell. Financial Eng. Econ. (CIFEr), Mar. 2014, pp. 114–121.

[17] Abimbola B, de La Cal Marin E, Tan Q. Enhancing Legal Sentiment Analysis: A Convolutional Neural Network–Long Short-Term Memory Document-Level Model[J]. Machine Learning and Knowledge Extraction, 2024, 6(2): 877-897.

[18] Gao Q, Wang D, Sun P, et al. Sentiment analysis based on the nursing notes on in-hospital 28-day mortality of sepsis patients utilizing the MIMIC-III database[J]. Computational and Mathematical Methods in Medicine, 2021, 2021.

[19] Waudby-Smith I E R, Tran N, Dubin J A, et al. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients[J]. PloS one, 2018, 13(6): e0198687.

Overflow

Literature Review

Categories	Models	Limitations	Mention Overlapping	Mix-polarity	No-result Cases
Clinical NLP software	CLAMP[4] cTAKES[5]	• Rely on pre-defined dictionary	X	-	-
ML methods	Support Vector Machine [6] Tree-based approach [7]	• Require complex feature engineering	X	-	-
DL methods	CBEx[8]: LSTM + dictionary-based exact and fuzzy match	• Fuzzy match introduces noise	✓	-	-
	[9] Pre-trained BERT model by clinical notes	• Not applicable when the data size is limited • Trained by non-English datasets	✓	X	X

Mining "Opinions" from Professional Documents

More complicated than mining public opinions from user-generated contents:

Longer document

Sentiment are domain-specific

Fine-grained sentiment analysis, ABSA, is necessary

E.g., The acquisition will considerably increase Kemira 's sales and market position.

Aspect	Polarity
Sales	Positive
Market position	Positive